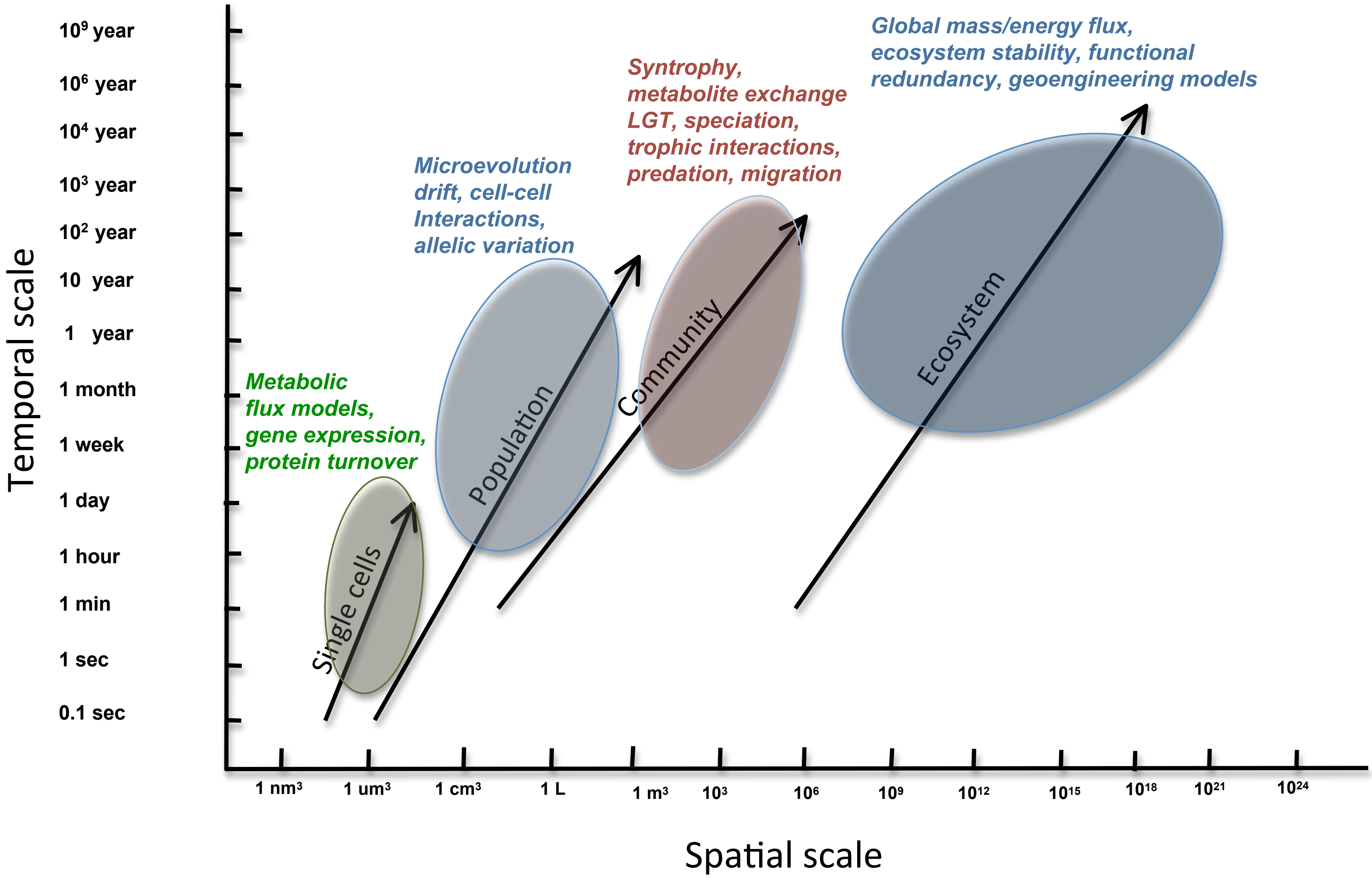
Ecological genomics and data drivers



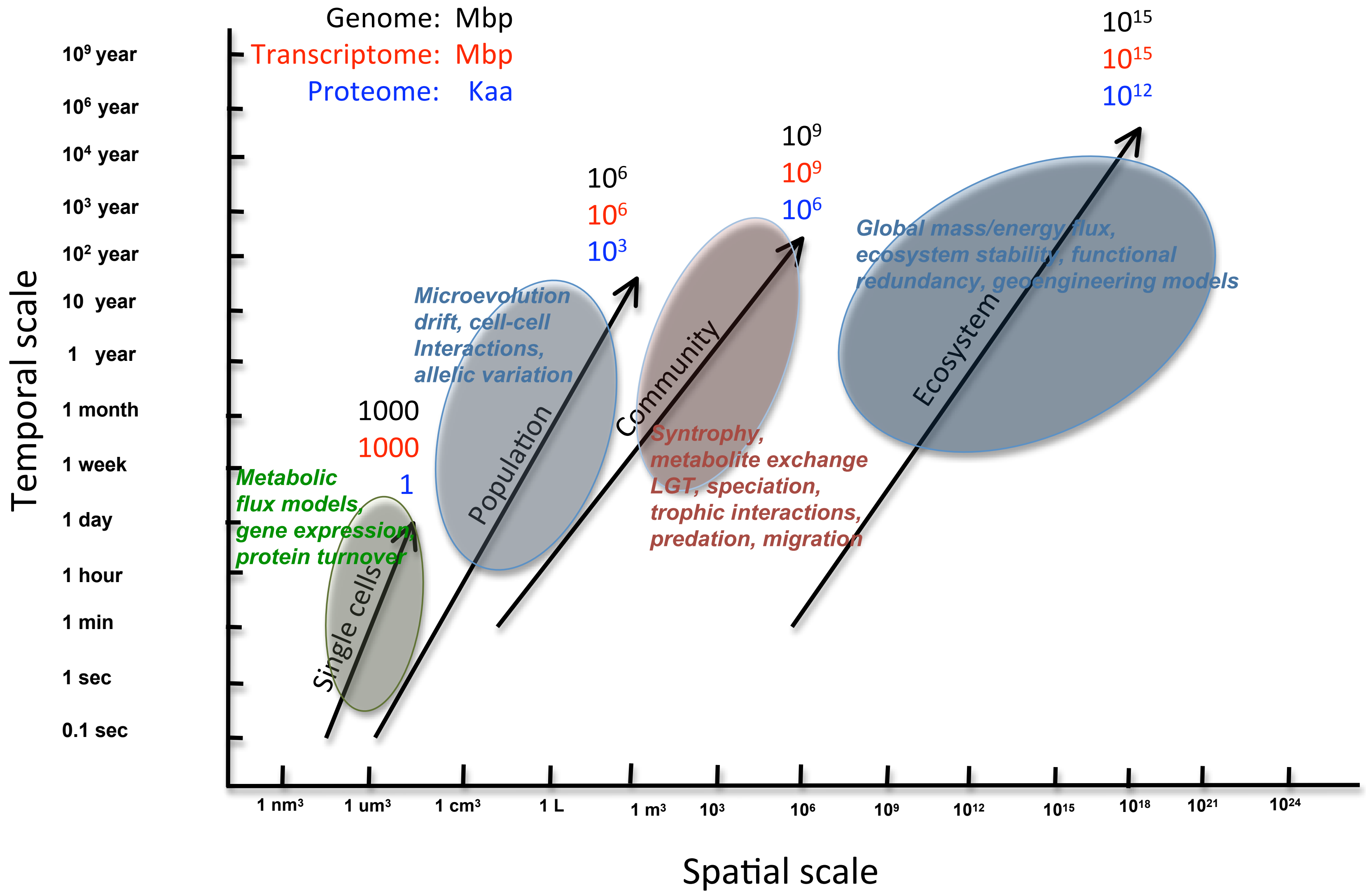
Metabolic flux models, gene expression, protein turnover

Microevolution drift, cell-cell Interactions, allelic variation

Syntrophy, metabolite exchange LGT, speciation, trophic interactions, predation, migration

Global mass/energy flux, ecosystem stability, functional redundancy, geoengineering models

Ecological genomics and data drivers



From Genomics & Populations to Ecosystems & Evolutionary Dynamics

Specific problem/challenge – what progress could be made with petaflops of computing power?

Grand Challenge for [microbial] ecology:

How the numerous taxa act and interact to assemble and sustain complex ecosystems.

Is the problem a "top 10" problems for the scientific discipline ? What's the community?

The top problem for ecology.

Community:

microbial ecology

anyone working in a microbial system

From Genomics & Populations to Ecosystems & Evolutionary Dynamics

Is petascale computational modeling irreplaceable in answering this question ?
How ?

Finally opens the door to realistically sized multiscale models

What is the current status of the computing tools for the work ?

Abstractions & mathematical models.

Black box concept of an organism (or group of organisms).

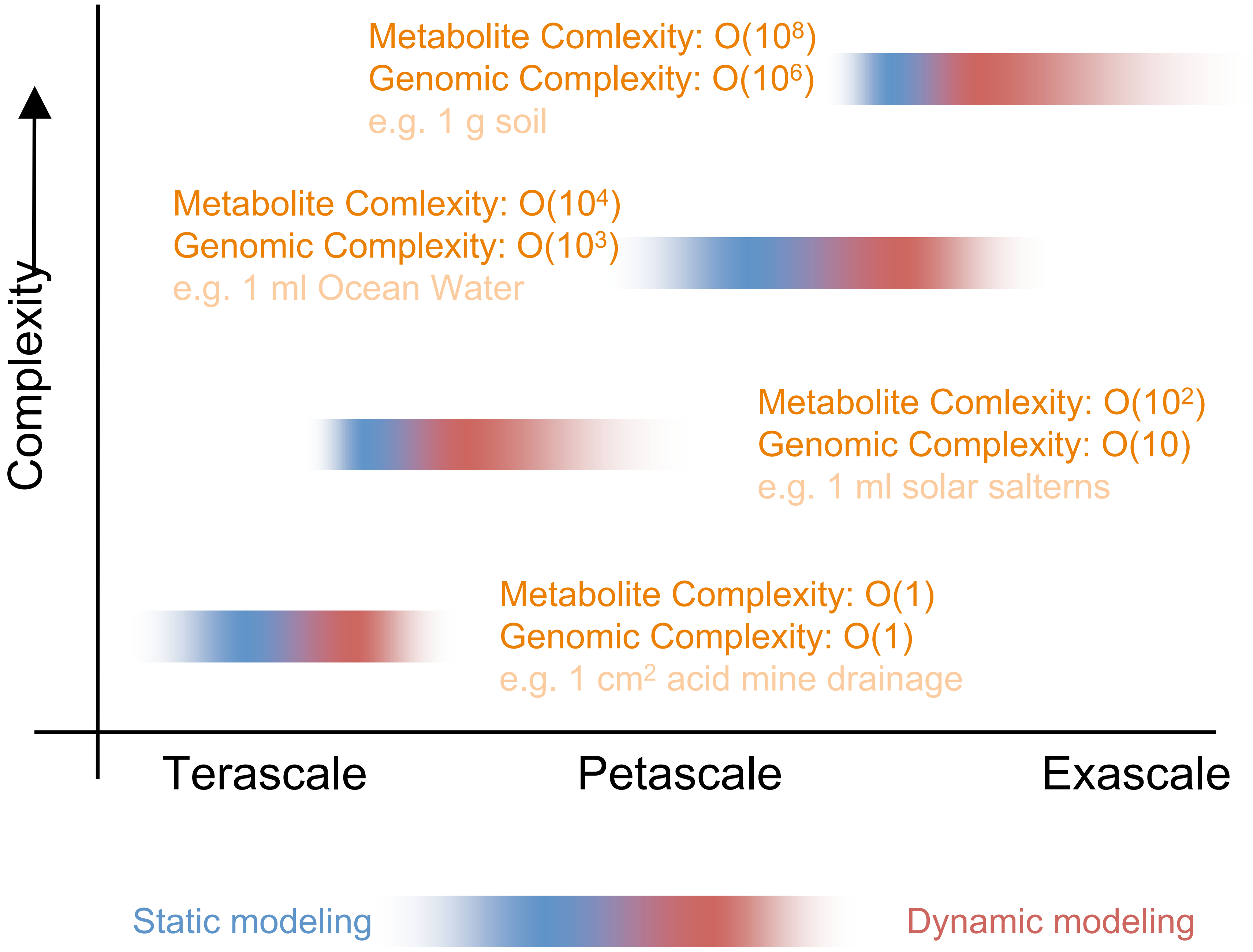
lacks metabolic detail; biological realism

What are the missing pieces ?
(mathematical models, algorithms, software, and data analysis tools?)

Dynamical systems models

Data fusion

Complex genotype/phenotype models


Metabolite Comlexity: $O(10^8)$
Genomic Complexity: $O(10^6)$
e.g. 1 g soil
Metabolite Comlexity: $O(10^4)$
Genomic Complexity: $O(10^3)$
e.g. 1 ml Ocean Water
Metabolite Comlexity: $O(10^2)$
Genomic Complexity: $O(10)$
e.g. 1 ml solar salterns
Metabolite Complexity: $O(1)$
Genomic Complexity: $O(1)$
e.g. 1 cm^2 acid mine drainage
Complexity
Terascale
Petascale
Exascale
Static modeling
Dynamic modeling

From Genomics & Populations to Ecosystems & Evolutionary Dynamics
-**Specific problem/challenge** what progress could be made with petaflops of computing power

The Challenge: Accurately model and engineer complex multispecies biological systems in both the lab and the environment. This challenge has as its goal the development of methods to measure, understand, simulate, model, and engineer complex, multispecies systems. It will involve the combination of the gathering of extensive meta-data along with extensive "omics" analyses, and the combination of these data sets with the goal of understanding carbon and energy flux through complex microbial communities. The studies will begin with fairly simple, defined laboratory mixed species systems, progressing in levels of complexity, and ending with the study of natural systems, utilizing similar methods and approaches.

Outputs/computational problem: The output of these efforts will be predictive, multi-scale simulation models of low to medium complexity microbial communities (complex models of simple communities). Problems

Technology engineering needs/enablers/drivers: This work will require improvements in annotation, with Google-like real time annotation/refreshing/recomputation across the protein universe. Methods for real-time metagenomics, metatranscriptomics, and metaproteomics are currently not available. In addition, technology for surveillance, high throughput metabolomics, and process monitoring will need to be developed. Few acceptable model systems have been descry bed, and will need to be developed. In addition, good methods for measuring and quantifying carbon and/or energy flux will be needed. Methods for correlating meta-data with omics data will need to be developed, and functional prediction in natural systems will need to be developed. Finally, there is little experience that can be used to guide the simulation work, and mathematical and computational tools are presently at a minimum.

From Genomics & Populations to Ecosystems & Evolutionary Dynamics

Is the problem a “top 10” problems for the scientific discipline ? What’s the community?

This problem is one of the big problems in environmental science – being able to predict and control the behavior and activities of microbial communities in natural systems would allow one to begin to make accurate estimates of carbon flow in natural environments of many sorts, and to be able to estimate the role of microbes in carbon uptake, sequestration, bioremediation, and other major processes. The community of users for this knowledge includes environmental scientists of all kinds.

From Genomics & Populations to Ecosystems & Evolutionary Dynamics

Is petascale computational modeling irreplaceable in answering this question ?

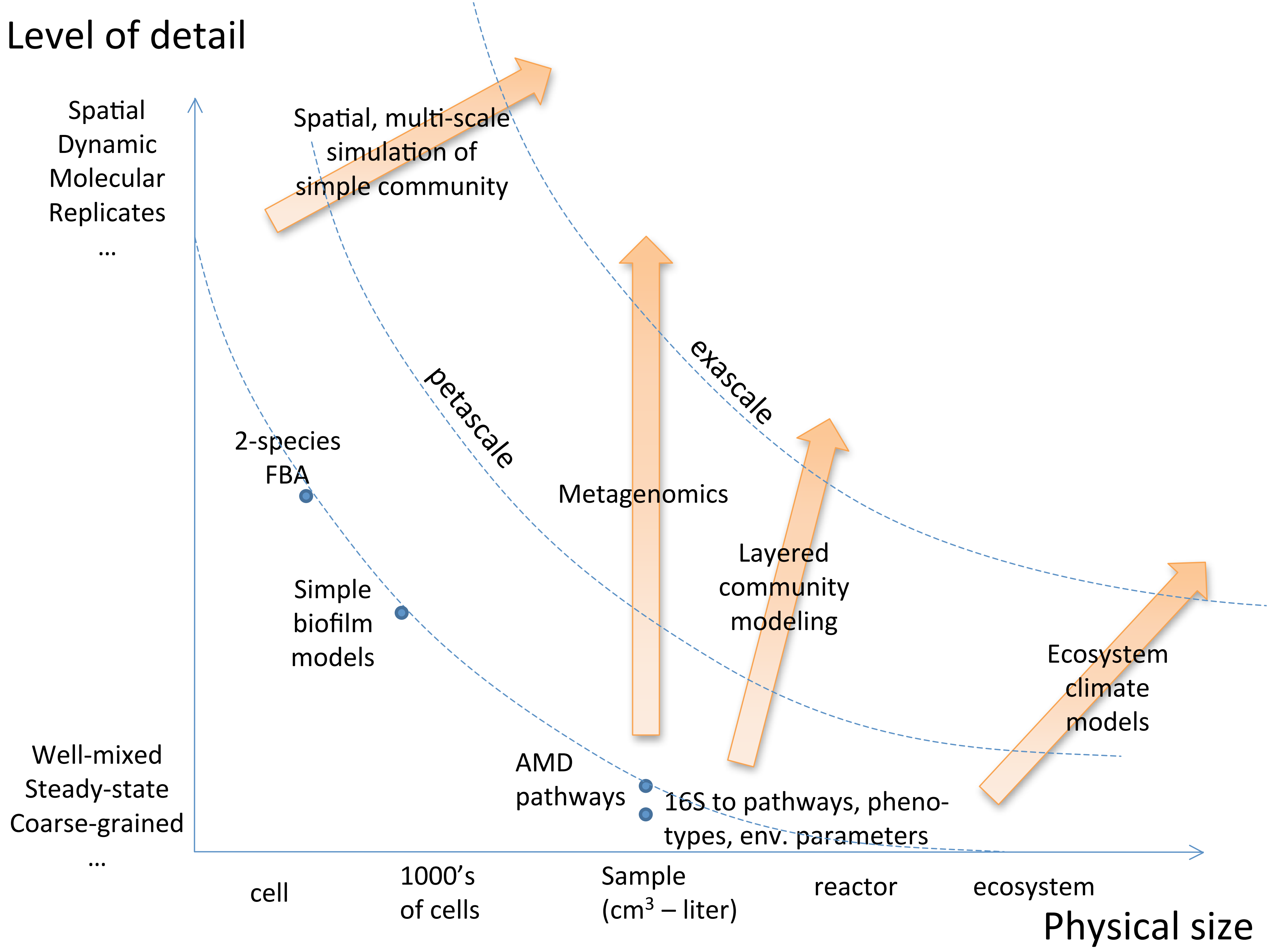
In its initial phases the work will be in the laboratory, and have rather limited computational needs, but as the program grows, and the systems become more complex, the need for petascale computing will arise, and the project will not be possible without it. Even the laboratory experiments, the chemostat or reactor – mixed culture systems used to predict dynamics in silico will become increasingly complex. It is this complexity that must be understood in order to simulate, model, predict, and engineer environmental communities, one of the true major challenges in biology today. Ultimately, we will establish an in-silico bioreactor with full spatio-temporal simulation capabilities (i.e., 3D-lattice simulations with diffusion dynamics, etc.). Multi-scale modeling and dynamic predictions made and tested with this system will be the precursor to environmental studies of even great complexity and computational needs. (see Figure).

What is the current status of the computing tools for the work ?

Several approaches for multiscale simulation of biological systems are emerging, including methods such as agent-based modeling, and other multiscale simulation approaches: e.g., coupled ordinary differential equations (ODE) and partial differential equations (PDE), kinetic Monte Carlo (KMC), and the cellular Potts model. However, to take advantage of the multiscale potential of exascale systems, further development of computational methods that enable coupling of intracellular, extracellular, and multicellular level reaction-diffusion model will be necessary.

What are the missing pieces ?
(mathematical models, algorithms, software, and data analysis tools?)

Further advances in computational methods to enable large scale computational biology will be necessary, including methods to couple data sets, to couple simulation methods, to adjust granularity and model adaptation with time and changes in conditions. Finally, simulation, methods for large-scale and advanced data visualization capabilities will be necessary.


Level of detail
Spatial
Dynamic
Molecular
Replicates
…
Well-mixed
Steady-state
Coarse-grained
…
Spatial, multi-scale
simulation of
simple community
petascale
exascale
Metagenomics
2-species
FBA
Simple
biofilm
models
Layered
community
modeling
Ecosystem
climate
models
AMD
pathways
16S to pathways, pheno-
types, env. parameters
cell
1000's
of cells
Sample
(cm^3 – liter)
reactor
ecosystem
Physical size

From Genomics & Populations to Ecosystems & Evolutionary Dynamics

Specific problem/challenge – what progress could be made with petaflops of computing power ?

- "Annotations" == function assignments to sequence features
- currently 30%-70% of annotations per genome are "**hypothetical**" or weak

1 2 3 4 5

- DNA → mRNA → Protein Sequence → Protein Fold → Docking of substrates

- current tools use sequence similarity of one or more proteins on level 1-3
- very little use of large scale fold (4) comparison and docking (5)

Is the problem a "top 10" problems for the scientific discipline ? What's the community?

- "Annotations" are a basic tool created by the genomics community and utilized by **many other communities** (microbiology, microbial ecology, metabolic modeling, …) for **many purposes** (studying cellulose degradation, carbon sequestration, N-cycling, remediation, ..)
- **Everyone** consuming genomic or metagenomic data will benefit
- This is a **Top 10 problem** in genome science.

Similar to Grand Challenge II of group
Macromolecular Proteins and Protein Complexes

From Genomics & Populations to Ecosystems & Evolutionary Dynamics

Is petascale computational modeling irreplaceable in answering this question ?
How ?

- current situation is an artifact from computability
- a lot of tools have already created the existing annotations, but rest is too hard
- **large scale folding and comparison of "hypotheticals"** could provide valuable guidance to guide experimentation

What is the current status of the computing tools for the work ?

- available tools for fold prediction, but require too many cycles
- automatic large scale comparison of many folds/shapes unclear

What are the missing pieces ?
(mathematical models, algorithms, software, and data analysis tools?)

- existing software is a start
- binning proteins to shape-classes **and** predicting binding and reaction mechanism